https://eventos.utfpr.edu.br//sicite/sicite2020

# Aprendizado de contexto de imagens por meio de redes neurais convolucionais baseadas em grafos

# Image context learning through graph convolutional networks

**Luis Gustavo de Souza**
souza.1998@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

**Priscila Tiemi Maeda Saito**
psaito@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

**Recebido:**
**Aprovado:**



## RESUMO

Entre as áreas da visão computacional, a compreensão de contexto de imagem, que se direciona ao reconhecimento de cenas pelos seus objetos, vem sendo observada em diferentes aplicações. Entretanto, ferramentas tradicionais da visão computacional, como as Redes Neurais Convolucionais (**CNNs**) e *Multi-Layer Perceptrons* (**MLPs**), por operarem no domínio euclidiano são incapazes de modelarem as relações entre os objetos da imagem, consequentemente, prejudicando o reconhecimento. Desta forma, este trabalho apresenta a proposta e o desenvolvimento de um modelo capaz de determinar o contexto de uma imagem pela análise do conteúdo dos objetos (características visuais) e as suas respectivas interações. Para isso, é construída uma representação em forma de grafo do conjunto de dados **MIT64,** e treinada uma estrutura de Rede Neural Convolucional baseada em Grafo (**GCN**) capaz de abstrair ambas informações (as interações e o conteúdo dos objetos). Pelos experimentos, a abordagem proposta apresenta, em relação às ferramentas tradicionais (**MLP**), uma melhora de 134% na assertividade do modelo. Assim, evidenciando a importância das interações entre os objetos em conjunto com as características visuais.

**PALAVRAS-CHAVE:** Aprendizado do computador. Processamento de imagens digitais. Visão por computador. Teoria dos grafos.

## ABSTRACT

Among the computer vision areas, the image context-reasoning, which is directed towards the recognition of the scenes by their objects, has been observed in different applications. However, traditional computer vision tools, such as Convolutional Neural Networks (**CNNs**) and Multi-Layer Perceptrons (**MLPs**), because operating in the Euclidean domain are unable to model the relationships between the objects of the image, consequently, impairing the recognition. Thus, this work presents the proposal and the development of a model capable of defining the image context by analyzing the content of the objects (visual features) and their respective interactions. For this purpose, a graph representation of the **MIT64** dataset is built, and a Graph Convolutional Network (**GCN**) structure is capable of reasoning both information (interactions and the content of objects). By the experiments, the proposed approach presents, in comparison with the traditional tools (**MLP**), an improvement of 134% in the accuracy of the model. Thus, highlighting the importance of the interactions between the objects in combination with the visual features.

**KEYWORDS:** Machine learning. Digital image processing. Machine vision. Theory of graphs.





## INTRODUÇÃO

Com a evolução dos dispositivos de *hardware* e *software,* o domínio da visão computacional emergiu-se, permitindo às maquinas a operação sobre tarefas complexas. Entre essas distintas tarefas, uma que vem sendo observada em diferentes aplicações, como na medicina (YOUNIS et al., 2019, p. 1630), agronomia (WANG et al., 2020, p. 105222) e varejo (BAZ; YORUK; CETIN, 2016, p. 1-5), trata-se da compreensão de contexto, que tem o objetivo de classificar/reconhecer determinada cena pelo seus objetos e suas respectivas interações.

Para esse problema, as Redes Neurais Convolucionais (**CNNs**) poderiam ser uma opção como ferramenta para reconhecer o contexto de determinado objeto. Entretanto, pelo fato de as **CNNs** operarem apenas no espaço euclidiano (BARRON, 1989, p. 280-285), essa ferramenta, naturalmente, é incapaz de estruturar as interações dos objetos, o que pode prejudicar o reconhecimento. Assim, de modo a superar as limitações desse gênero, a literatura, por meio da capacidade dos grafos na modelagem de dados não-euclidianos e com o poder de generalização das **CNNs**, tem proposto abordagens baseadas em Redes Neurais Convolucionais baseadas em Grafos (**GCNs**).

Entre as estruturas de **GCNs** disponíveis, pode-se destacar a **GCN** espectral de Kipf e Welling (2016), que é fundamentada no processamento de sinais de grafos, e por conta disso, representa os nós (objetos das imagens) do grafo como sinais e agregam as informações da vizinhança por meio da transformada de *Fourier*. Sendo que, para essa estrutura, a saída de dados é dada pela classificação de cada nó do grafo (no caso deste trabalho, como demonstrado na seção seguinte, é a classe de contexto global do objeto em questão).

As aplicações das **GCNs** são dedicadas, geralmente, às redes de citação para classificação de documentos, negligenciando a área de visão computacional. Desta forma, baseado em Bugatti, Saito e Davis (2019), este trabalho tem por objetivo construir um modelo, utilizando **GCNs** (KIPF; WELLING, 2016), capazes de compreender o contexto de imagens pelas características visuais dos objetos e as suas interações.

Especificamente, as direções deste trabalho são: construir um grafo completo, independente de grafo de conhecimento e da classe individual de cada objeto; compreender o contexto de imagens de ambientes fechados por meio de uma estrutura de **GCN**; comparar as abordagens tradicionais (MLP) com a proposta (GCN) na modelagem de compreensão de contexto.

## MATERIAIS E MÉTODOS

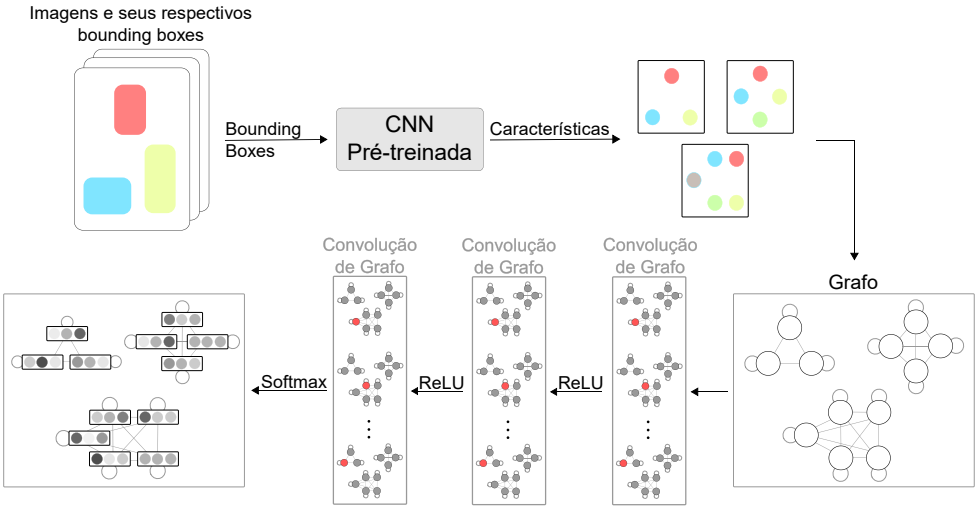
A essência deste estudo, como demonstrado pela Figura 1, remete-se à construção de um modelo capaz de definir a classe de contexto de um objeto, utilizando as relações semânticas com outros objetos associados. Com essa direção, o conjunto de dados é representado por meio de um grafo não direcional, que como discutido previamente, é adequada para esse tipo de modelagem.

Para a construção do grafo **G**, os objetos das imagens (*bounding boxes*) de potencial discriminação, ou seja, que possuem informações relevantes para a determinação do contexto da imagem, são coletados de forma a representar o

conjunto de nós **V** do grafo. O conjunto de arestas **E** determina quais bounding boxes/nós estão interligados entre si.

Contudo, para o processo de treinamento, o conteúdo dos nós precisam ser representados em vetores numéricos. Assim, uma **CNN** pré-treinada é utilizada para extrair as características visuais de cada *bounding box* para a geração dos vetores. Como procedimento padrão, os *bounding boxes* são remodelados, podendo alterar sua proporção, de forma a atingir um tamanho de 224 pixels x 224 pixels x 3 canais de cores. Note que, outras características poderiam ser avaliadas, como: as espaciais. Entretanto, essa característica foi desconsiderada neste estudo, pois como demonstrado em Bugatti, Saito e Davis (2019), sua adição apresentou um impacto insignificante na inferência do modelo.

Figura 1 – As etapas do processo proposto




Fonte: Autoria própria (2020).

Em uma configuração supervisionada, os estudos (WANG et al., 2018, p. 1021–1028)(YANG et al., 2019, p. 1–13), geralmente, exigem conhecer os rótulos de cada amostra (neste caso, dos *bounding boxes*/nós), o que pode ser inviável para conjuntos de dados com diversos objetos distintos. Desta forma, baseado em Bugatti, Saito e Davis (2019), este trabalho concede aos *bounding boxes* a classe da imagem global. Por exemplo, em uma imagem de aeroporto existe uma pessoa (*bounding box*) que receberá o rótulo **aeroporto** ao invés de **pessoa**.

Outro recurso exigido por alguns trabalhos (CHEN et al., 2018, p. 7239-7248) (CHEN; SHRIVASTAVA; GUPTA, 2013, p. 1409-1416)(MARINO; SALAKHUTDINOV; GUPTA, 2016), trata-se do grafo de conhecimento, que define previamente as relações entre os nós. No entanto, os grafos de conhecimento são, em sua maioria, custosos e escassos para serem obtidos. Em vista disso, este trabalho constrói um grafo completo entre os *bounding boxes* de uma mesma imagem (ou seja, todos os *bounding boxes* de uma imagem são interligados entre si), e acrescenta auto conexões (o nó conecta-se a si mesmo).

Finalmente, com o grafo construído, um modelo de **GCN**, definido por Kipf e Welling (2016), é criado para realizar o treinamento sobre todas as amostras rotuladas, por meio da função custo *cross-entropy* utilizando a técnica de retro propagação com o otimizador **ADAM**. Com isso, ao final do treinamento espera-se

que o modelo gere representações poderosas para os nós, de forma a classificar o seu contexto adequadamente.

## RESULTADOS E DISCUSSÃO

Nesta seção é discutida a especificação do conjunto de dados, evidenciando as suas estatísticas em conjunto com o grafo desenvolvido. Em seguida, são demonstrados os experimentos realizados na área de compreensão de contexto, com a utilização do modelo proposto (**GCN**) e tradicional (**MLP**).

Para os experimentos, o conjunto de dados **MIT64** (QUATTONI; TORRALBA, 2009, p. 413-420) foi escolhido, pois fornece os requisitos necessários para o desenvolvimento deste trabalho, sendo: as imagens, os *bounding boxes* de cada imagem, e as classes globais (classes das imagens). O **MIT64** trata-se de um conjunto de dados para solucionar o reconhecimento de ambientes fechados, contendo 67 classes distintas. Entretanto, similar a Bugatti, Saito e Davis (2019), pelo fato de algumas classes possuírem poucos exemplos, erros de anotação, dados faltantes ou imagens sem bounding boxes, uma limpeza dos dados foi demandada, o que resultou na exclusão das seguintes classes: *auditorium, bowling, elevator, jewellery shop, locker room, hospital room, restaurant kitchen, subway, laboratory wet, movie theater, museum, nursery, operating room, waiting room*. Assim, obtendo 53 classes, 2607 imagens e 50.868 bounding boxes, que por sua vez, foram divididos nos conjuntos de treino (80% dos dados) e teste (20% dos dados), de forma aleatória e estratificada.

Com o conjunto de dados transformado e na intenção de compreender a sua estrutura, gerou-se a Tabela 1, que demonstra as informações gerais do conjunto de dados. Pode-se perceber pelo desvio padrão da tabela, um desbalanceamento acentuado entre as classes, o que pode impactar no treinamento.

Tabela 1 – Estatísticas do conjunto de dados **MIT64**

| Análise | Média | Desvio P. | Mínimo | Mediana | Máximo |
|---|---|---|---|---|---|
| Imagens por classe | 49,2 | 68,1 | 16 | 19 | 350 |
| *Bounding boxes* por classe | 959,8 | 1311,7 | 159 | 475 | 7511 |
| *Bounding boxes* por imagem | 18,8 | 12,7 | 1 | 17 | 95 |

Fonte: Autoria própria (2020).

Para o treinamento, primeiramente, foram considerados 4 diferentes arquiteturas CNNs para a extração das características: **VGG19** (512 características), **ResNet50** (2048), **InceptionV3** (2048) e **EfficientNetB7** (2560). Em seguida, 2 camadas de convoluções de grafo (modelo **GCN**) ou densas (modelo **MLP**), ativadas por meio da função **ReLU,** são adicionadas. Por fim, uma camada de saída com ativação *softmax* é situada no topo de ambas estruturas, de modo a obter as probabilidades de cada classe.

Com essas estruturas, treinou-se ambos modelos para cada combinação dos seguintes hiper parâmetros: *epochs* (2000), *hidden units* (32, 128, 256), *learning rate* ($1e^{-2}$, $5e^{-2}$, $1e^{-3}$, $5e^{-3}$, $1e^{-4}$, $5e^{-4}$), *dropout* (0.3, 0.5, 0.8, 0.9). Ao fim, foi coletado

o melhor modelo para inferir sobre o conjunto de teste, obtendo as métricas constatadas na Tabela 2.

De acordo com os resultados da Tabela 2, o modelo proposto, considerando cada arquitetura (**VGG19**, **ResNet50**, **InceptionV3** e **EfficientNetB7**), supera significativamente o modelo tradicional. Isso, evidencia a importância das interações entre os objetos, que o modelo proposto consegue modelar, diferentemente dos modelos tradicionais.

Tabela 2 – Resultados obtidos pelas abordagens **GCN** (proposta) e **MLP** (tradicional)

| Estrutura | Extrator | Acurácia (%) | Recall (%) | Precisão (%) | F1 (%) |
|---|---|---|---|---|---|
| **GCN** | VGG19 | 60,62 | 46,68 | 49,87 | 45,97 |
| | ResNet50 | 70,56 | 59,27 | 59,94 | 57,35 |
| | InceptionV3 | **75,23** | **66,46** | 66,84 | 64,92 |
| | EfficientNetB7 | 75,19 | 66,28 | **71,65** | **65,69** |
| MLP | VGG19 | 27,83 | 17,40 | 35,73 | 20,77 |
| | ResNet50 | 30,36 | 20,89 | 32,43 | 23,38 |
| | InceptionV3 | 28,24 | 16,86 | 34,11 | 19,47 |
| | EfficientNetB7 | 32,06 | 21,88 | 34,97 | 24,46 |

Fonte: Autoria própria (2020).

## CONCLUSÃO

Como visto, foi proposta uma abordagem para suprir a área de compreensão de contexto, utilizando as redes neurais convolucionais baseadas em grafos capazes de modelarem as características visuais e as interações dos objetos da imagem. Além disso, diferentemente dos trabalhos que utilizam os grafos em visão computacional, este trabalho faz o uso mínimo de recursos, esquivando-se do grafo de conhecimento e das classes individuais de cada *bounding box*.

Se considerado o melhor cenário para ambos os modelos, o modelo proposto obteve uma acurácia 134,65% superior ao modelo tradicional, evidenciando o benefício de se estruturar as interações entre os objetos para o problema de compreensão de contexto.

Para os trabalhos futuros, pretende-se tornar o modelo ainda mais independente pela identificação automática de *bounding boxes*, de modo a treiná-lo fornecendo apenas as imagens e as suas respectivas classes. Além disso, fazer o uso de pesos sobre as arestas, que podem aumentar o desempenho do modelo. Uma última direção, que poderia ser explorada, refere-se à generalização desse modelo para diferentes áreas de visão computacional.

## AGRADECIMENTOS

O presente estudo foi realizado com o apoio das instituições: Universidade Tecnológica Federal do Paraná e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Processo 431668/2016-7.

## REFERÊNCIAS

BARRON, A R. **Statistical properties of artificial neural networks**. *In*: Proceedings of the 28th IEEE Conference on Decision and Control. IEEE, 1989. p. 280-285.

BAZ, I.; YORUK, E.; CETIN, M. **Context-aware hybrid classification system for fine-grained retail product recognition**. *In*: 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). IEEE, 2016. p. 1-5.

BUGATTI, P. H.; SAITO, P.; DAVIS, L. S. **HiCoRe**: Visual Hierarchical Context-Reasoning. arXiv preprint arXiv:1909.00848, 2019.

CHEN, X. et al. **Iterative visual reasoning beyond convolutions**. *In*: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 7239-7248.

CHEN, X.; SHRIVASTAVA, A.; GUPTA, A. **Neil**: Extracting visual knowledge from web data. *In*: Proceedings of the IEEE international conference on computer vision. 2013. p. 1409-1416.

KIPF, T. N.; WELLING, M. **Semi-supervised classification with graph convolutional networks**. arXiv preprint arXiv:1609.02907, 2016.

MARINO, K.; SALAKHUTDINOV, R.; GUPTA, A. **The more you know**: Using knowledge graphs for image classification. arXiv preprint arXiv:1612.04844, 2016.

QUATTONI, A.; TORRALBA, A. **Recognizing indoor scenes**. *In*: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009. p. 413-420.

WANG, F. et al. **Fusing multi-scale context-aware information representation for automatic in-field pest detection and recognition**. Computers and Electronics in Agriculture, 2020. v. 169, p. 105222.

WANG, Z. et al. **Deep reasoning with knowledge graph for social relationship understanding**. *In:* International Joint Conferences on Artificial Intelligence, 2018. p. 1021–1028.

YANG, W. et al. **Visual semantic navigation using scene priors**. *In*: International Conference on Learning Representations, 2019. p. 1–13.

YOUNIS, O. et al. **A smart context-aware hazard attention system to help people with peripheral vision loss**. Sensors, 2019. v. 19, n. 7, p. 1630.